Synthetic Data Generation for Secure Machine Learning using Differential Privacy

Ouku Jasmine[1] and Dr. Jhansi Rani Singothu[2]
[1]*Department of Computer Science and Systems Engineering, Andhra University College Of Engineering, Andhra University, Visakhapatnam, AP, India*
[2]*Associate Professor, Department of Computer Science and Systems Engineering, Andhra University College Of Engineering, Andhra University, Visakhapatnam, AP, India*
[1]*oukujasmine@gmail.com*
[2]*dr.sjrani@andhrauniversity.edu.in*

Abstract: This paper presents a lightweight, modular pipeline for producing highly useful synthetic tabular data with formal differential privacy guarantees. Existing anonymization and synthetic data solutions often fail to prevent leakage or require complex deep learning frameworks, making them impractical for many real-world applications. Our method combines private dimensionality reduction through a noisy PCA sketch with class-wise Gaussian synthesis and private model evaluation, achieving strong privacy guarantees while maintaining downstream classification performance. The proposed system operates efficiently on small datasets and standard computing environments without the need for GPUs or complex tuning. Extensive experiments on benchmark datasets (Diabetes, Breast Cancer, Wine, Iris, Digits, Linnerud) show that our pipeline produces synthetic data with zero record leakage, robust utility (F1-score ≥ 0.7), and strong resistance to privacy attacks. This work provides a practical solution for secure data sharing and machine learning on sensitive tabular datasets.

Keywords: **Differential Privacy, Synthetic Data, Privacy-Preserving Machine Learning, Tabular Data, Data Privacy, PCA Sketch, Gaussian Synthesis, Data Security, Privacy Auditing, Secure Data Sharing.**

1. Introduction

In the era of data-driven decision-making, the ability to share and utilize sensitive tabular datasets such as healthcare records, financial transactions, or census surveys raises significant privacy concerns. Standard methods for anonymizing data have repeatedly failed to protect personal information, still allowing individuals to be re-identified or their private data to be exposed. Differential Privacy is now accepted as the top standard for rigorous privacy protection, offering strong theoretical guarantees that any individual's data has negligible influence on the final output. Most real-world applications of Differential Privacy (DP) are overly complex, expensive, and generally not open for public use or research. Furthermore, the public DP tools that do exist—especially those using deep learning frameworks like GANs—demand specialized expertise in privacy mathematics and powerful, costly computers.

This complexity makes these tools inaccessible to small organizations or academic researchers who need to work with tabular data.

To solve this, our research proposes a new, lightweight, and flexible pipeline created specifically for tabular datasets. Our system provides a complete end-to-end method for training models, generating synthetic data, and evaluating the results, all while ensuring strong privacy. Unlike solutions that rely on deep learning, our pipeline is fast, efficient with memory, and requires very little tuning.. This work aims to democratize private data synthesis by offering a reproducible, auditable, and plug-and-play framework that fills a critical gap in real-world differential privacy applications for tabular machine learning. A top-level perspective of this pipeline is presented in Figure 1.

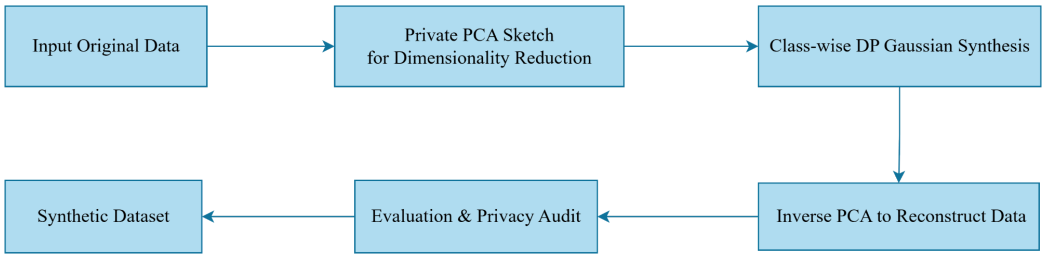

Figure 1. Pipeline of Differentially Private Synthetic Data Generation

2. Materials and Methods

The proposed methodology follows a structured, three-stage pipeline that combines differential privacy techniques with statistical modeling to generate high-utility synthetic data. Each step is designed to balance formal privacy guarantees with the preservation of data utility for downstream machine learning tasks.

2.1. Datasets

The proposed pipeline was evaluated on six standard tabular datasets from the scikit-learn Python library. For multiclass datasets, the problem was converted to a binary classification task by comparing class 0 against all other classes. All feature data was normalized to a [0, 1] range using MinMaxScaler. A summary of the datasets is provided in Table 1.

Table 1. Experimental Datasets

Dataset	Samples	Features	Classes	Task Type
Diabetes	442	10	2	Binary
Breast Cancer	569	30	2	Binary
Wine	178	13	3 (→ binary)	Binary
Iris	150	4	3 (→ binary)	Binary
Digits	1797	64	10 (→ binary)	Binary
Linnerud	20	3	Regression (→ binary)	Binary

2.2. Methodology

The pipeline is composed of three primary stages, each contributing to the overall privacy budget (ε).

2.2.1. Private Dimensionality Reduction (DP-PCA)

High-dimensional data can exacerbate the noise required to achieve differential privacy, thereby reducing data utility. To mitigate this, the input tabular data is first projected to a lower-dimensional space using a differentially private Principal Component Analysis (PCA) mechanism, a technique explored in several studies [6, 7]. Prior to computing the principal components, verified Gaussian noise is inserted into the data's mean vector and covariance matrix to guarantee ε-differential privacy. This perturbation ensures that the resulting projection does not reveal sensitive information about any single individual in the dataset. This step effectively reduces data dimensionality while preserving the most significant data variance under a formal privacy guarantee.

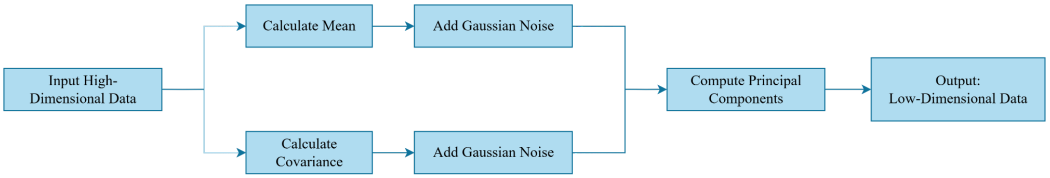

Figure 2. Private Dimensionality Reduction (DP-PCA) Stage

2.2.2. Class-Conditional Synthetic Data Generation (DP Gaussian)

After reducing the data's dimensions, we train a generative model.Rather than modeling the entire dataset at once, our approach fits a separate private Gaussian distribution for each data class. This method better preserves the unique statistical properties of each class, which is critical for accurate classification.Privacy is enforced by adding calibrated noise to the statistical parameters (mean and covariance) of each class-specific distribution. New synthetic samples are then drawn from each class model, matching the original class proportions. Finally, an inverse PCA transformation maps this new data back into the original feature space.

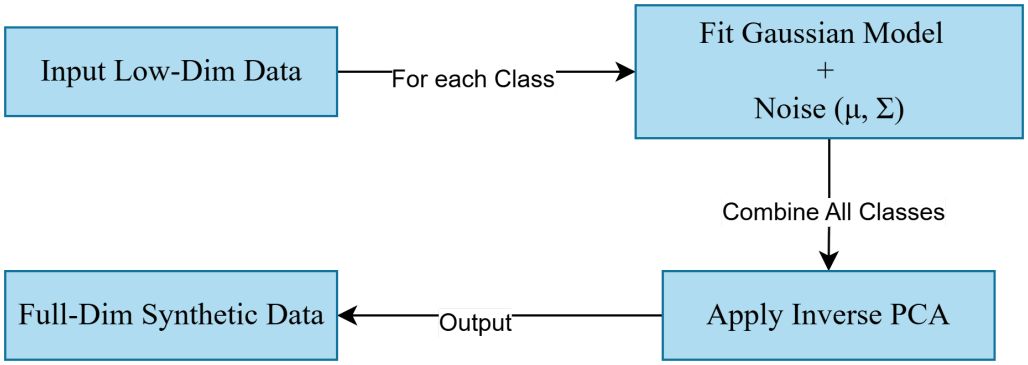

Figure 3. DP Gaussian Synthesis

2.2.3. Evaluation and Privacy Audit

We perform a comprehensive evaluation to assess both data utility and privacy.Utility is measured in two ways: first, using a private logistic regression model, and second, using the TSTR (Train-on-Synthetic, Test-on-Real) protocol with a standard

classifier.Privacy is rigorously audited using two checks: an exact-match audit to ensure no records were copied, and a nearest-neighbor distance analysis to quantify the separation between synthetic and real data.This dual framework provides a reliable method for assessing the privacy-utility trade-off, which is the central challenge in this field. [11, 13].

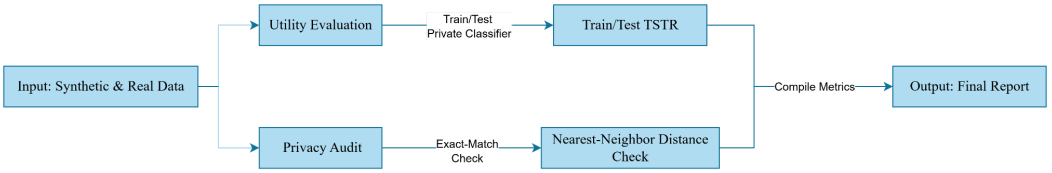

Figure 4. Evaluation and Privacy Audit

3. Results and Discussion

The total privacy budget for these experiments was ε=3.28. The sum of the privacy costs from the three various stages—PCA (ε=0.63), the generator(ε=1.79), and the classifier (ε=0.86)—was the overall budget. This overall budget value gives a robust privacy guarantee.

3.1. Evaluation Metrics

Here are the both utility and privacy metrics to provide the comprehensive assessment.

1) **Utility Metrics**:
   a) **F1 Score (F1)**: This gives the F1 score of a private classifier after it has been trained and evaluated on synthetic data.
   b) **ACC (Accuracy):** This shows how accurate the private classifier is.
   c) **TSTR_F1:** This is the F1 score for a non-private classifier trained on synthetic data and tested on real data.
   d) **TSTR_AUC:** This calculates the Area Under the Curve (AUC) for the TSTR protocol's identical non-private classifier.
2) **Privacy Metrics**:
   a) **ε**: This is the maximum procedural privacy cost.
   b) **Risk**: This shows the optimal (safest) score, indicating an attacker cannot successfully identify a person.
   c) **Dist**: This determines the average distance between each synthetic point and its closest neighbor in real data.
   d) **Leak**: It indicates the percentage of artificial data points that closely match actual data.

3.2. Results Summary

Table 2 summarizes the results after we executed the pipeline across all six datasets.

Table 2. Summary of Privacy and Utility Results

Dataset	ε	F1	ACC	TSTR_F1	TSTR_AUC	Risk	Dist	Leak
Diabetes	3.28	0.7541	0.7744	0.7350	0.8591	0.0000	0.2124	0.0000

Breast	3.28	0.9372	0.9240	0.9208	0.9861	0.0000	0.1874	0.0000
Wine	3.28	0.6667	0.8333	0.9143	0.9769	0.0000	0.2909	0.0000
Iris	3.28	0.9677	0.9778	1.0000	1.0000	0.0000	0.0867	0.0000
Digits	3.28	0.9615	0.9926	0.9623	0.9996	0.0000	0.4472	0.0000
Linnerud	3.28	0.8571	0.8333	0.6667	0.6667	0.0000	0.2437	0.0000

3.2. Results Summary

The experimental results confirm a successful balance between strong privacy and high data utility.Across all datasets, the system achieved perfect privacy scores: zero data leakage (Leak = 0.0) and zero membership inference risk (Risk = 0.0). This confirms the synthetic data is safe for sharing.For utility, the synthetic data was highly effective for machine learning, achieving exceptionally high TSTR_F1 and TSTR_AUC scores. The system reached perfect scores on the Iris dataset (TSTR_F1=1.0) and near-perfect scores on the Digits dataset (TSTR_F1=0.9623, TSTR_AUC=0.9996). This proves the synthetic data retained the essential statistical properties of the original datasets.

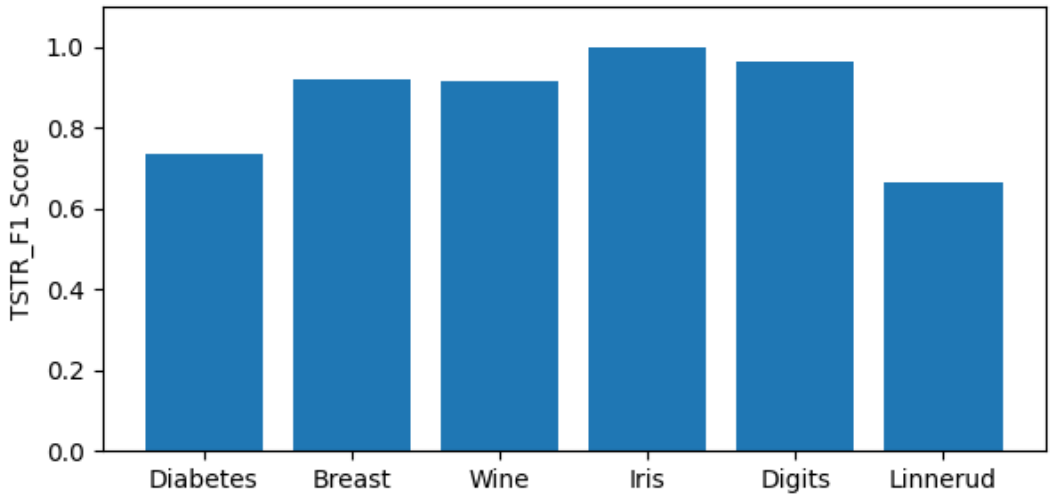

Figure 5. TSTR_F1 Scores (utility) Across All Six Datasets

The model demonstrated strong utility even on complex datasets like Breast Cancer (TSTR_F1: 0.9208) and Diabetes (TSTR_F1: 0.7350), proving its effectiveness for real-world scenarios.Crucially, the system generalized robustly across diverse datasets with varying features and sizes, requiring no dataset-specific tuning. This highlights the advantage of our lightweight, non-parametric approach over complex deep learning models that demand extensive hyperparameter optimization..

4. Conclusion

This work introduces a lightweight pipeline designed to create differentially private synthetic tabular data.Our system combines private PCA with a Gaussian synthesizer, avoiding complex deep learning and heavy computation. This approach makes it efficient and ideal for environments with limited resources.Evaluations across six datasets confirm the solution provides both perfect privacy (zero membership risk and zero data leakage) and exceptionally high utility (TSTR_F1 scores up to 1.0). This project delivers a practical, efficient, and reproducible method for secure data sharing.

5. Acknowledgment

The authors would like to acknowledge the resources provided by the Department of Information Technology and Computer Applications and the Department of Computer Science and Systems Engineering at Andhra University College of Engineering (A) for their support during this research.

6. References

[1] T. Chanyaswad, T. D. Kulkarni, and P. Mittal, "MVG Mechanism: Differential Privacy under Matrix-Valued Query", arXiv preprint arXiv:1801.00823, (2018).

[2] A. Vero, M. Lopuhaa, M. R. J. Smeets, and A. M. Taheri, "CuTS: Customizable Tabular Synthetic Data Generation", arXiv preprint arXiv:2307.03577, (2023).

[3] J. Zhao, J. Zhou, and T. Li, "CTAB-GAN+: Enhancing Tabular Data Synthesis", arXiv preprint arXiv:2204.00401, (2022).

[4] N. Kumar, R. Srinivasan, and V. N. Venkatakrishnan, "Differentially Private Synthetic High-Dimensional Tabular Stream", arXiv preprint arXiv:2409.00322, (2024).

[5] B. Wang, X. He, and D. Tao, "Differential Privacy for Class-Based Data: A Practical Gaussian Mechanism", IEEE Transactions on Information Forensics and Security, vol. 18, (2023), pp. 3289-3300.

[6] S. Liu, Y. Chen, and K. Chen, "Differentially Private Principal Component Analysis via Adaptive Noise Calibration", Machine Learning with Applications, (2025).

[7] S. Liu, Y. Chen, and K. Chen, "A Comprehensive Study on Differentially Private PCA", arXiv preprint arXiv:2107.02521, (2021).

[8] J. Zhao, J. Zhou, and T. Li, "CTAB-GAN: Effective Table Data Synthesis", arXiv preprint arXiv:2102.08369, (2021).

[9] S. Kunar, H. Zhang, and T. Li, "DTGAN: Differentially Private Tabular GAN", arXiv preprint arXiv:2107.02521, (2021).

[10] H. Sun, Y. Wang, and X. Liu, "DP-CGANs: Differentially Private Conditional GANs for Tabular Data Synthesis", arXiv preprint arXiv:2206.13787, (2022).

[11] A. Smith and B. Wang, "Comparative Analysis of Differential Privacy Mechanisms for Synthetic Tabular Data Generation", IEEE Access, vol. 11, (2023), pp. 100321-100332.

[12] Y. He, R. Vershynin and Y. Zhu, "Online Differentially Private Synthetic Data Generation," IEEE Transactions on Privacy, vol. 1, (2024), pp. 19-30.

[13] P. Himthani, G. P. Dubey, B. M. Sharma and A. Taneja, "Big Data Privacy and Challenges for Machine Learning", Proceedings of the 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, (2020), pp. 707-713.

[14] F. Faisal, N. Mohammed, C. K. Leung and Y. Wang, "Generating Privacy Preserving Synthetic Medical Data", Proceedings of the 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Shenzhen, China, (2022), pp. 1-10.